Problemas de clasificación: problemas de localización

Emilio Carrizosa
Facultad de Matemáticas
Universidad de Sevilla
ecarrizosa@us.es

Belén Martín-Barragán
Facultad de Matemáticas
Universidad de Sevilla
belmart@us.es

Resumen

En estas páginas se describen algunas conexiones entre algunos métodos de Clasificación Supervisada (Máquinas de Vector de Apoyo, Vecinos Más Cercanos) y ciertos problemas de Localización.

1. Introducción

En la última década, la capacidad de almacenamiento de información digital se ha duplicado cada nueve meses. Crece, por tanto, a una velocidad muy superior a la prevista por la ley de Moore para el crecimiento de la capacidad de cálculo, [12, 16], provocando la aparición de las denominadas *fosas de datos*, [12]: datos que son almacenanados y descansan en paz, sin que nadie los reclame o los recuerde.

Con este panorama se ha desarrollado una disciplina conocida como *Minería de Datos*, cuyo objetivo es explorar grandes volúmenes de datos para extraer información previamente inesperada y potencialmente útil.

Usando herramientas matemáticas de diversos campos, como la Investigación Operativa, la Estadística o la Inteligencia Artificial, está siendo aplicada en multitud de sectores, donde se generan bases de datos de gran tamaño, en ocasiones muy desectruturadas y con ruido (valores perdidos, valores erróneos, ...). Entre ellos podríamos citar la Bioinformática (expresión genética, ...), el Márketing (segmentación de clientes, medida del éxito de campañas,...), Gestión de Empresas (Customer Relation Management,...), la Banca (valoración de riesgo en créditos a clientes, detección de fraude en tarjetas de crédito ...), Internet (Minería de páginas web), o la Farmacología (diseño de medicamentos, ...). Las referencias [1, 3, 13, 14, 17] pueden servir de introducción al tema.

Uno de los cometidos básicos de la Minería de Datos es *clasificar* casos, a través de la llamada *clasificación supervisada*. Tenemos un conjunto de objetos Ω. Cada objeto $u \in \Omega$ tiene dos componentes $u = (x^u, c^u)$, donde $x^u \in \mathcal{X}$ (usualmente $x^u \in \Re^p$) representa el vector de variables predictoras, y $c^u \in \mathcal{C}$ es la clase a la que pertenece u. Se dispone de un conjunto no vacío de objetos $I \subset \Omega$, la *muestra de aprendizaje*. El objetivo es predecir, a partir de I, la clase c^v a la que pertenece un objeto $v \in \Omega$ conociendo sólo x^v. Por ejemplo, en una aplicación de diagnóstico médico, se usarán los datos de un conjunto de pacientes de los que sabemos con certeza si tienen o no una cierta enfermedad como muestra de aprendizaje, y con esta información se construirá la regla, que será usada para diagnosticar a futuros pacientes.

Describimos a continuación dos estrategias de clasificación, estrechamente relacionadas con ciertos problemas de localización: las Máquinas de Vector de Apoyo y los Métodos de Vecino Más Cercano.

2. Máquinas de Vector de Apoyo

Por simplicidad en la exposición, supondremos el caso binario, $\mathcal{C} = \{-1, 1\}$, y que $\mathcal{X} = \Re^p$. Se buscan $\omega \in \Re^p$, $\beta \in \Re$, se construye la función de evaluación f,

$$f(x) = \omega^\top x + \beta, \tag{1}$$

y con ésta, la *regla lineal de clasificación* que clasifica en el grupo 1 a aquellos $x \in \Re^p$ con $f(x) > 0$ y en el grupo -1 a los x con $f(x) < 0$. Los x con $f(x) = 0$ serán clasificados siguiendo alguna regla predeterminada.

La primera pregunta que nos hacemos es si existen o no ω, β tales que la correspondiente regla lineal clasifique correctamente el 100 % de los individuos de I,

$$y^u \left(\omega^\top x^u + \beta\right) > 0 \quad \forall u \in I. \tag{2}$$

2.1. El caso separable.

Cuando el sistema (2) sea factible, diremos que I es *separable linealmente.* Cualquier (ω, β) solución de (2) satisface que $\omega \neq 0$. En particular, (ω, β) genera un hiperplano, $\{x \in \Re^p : \omega^\top x + \beta = 0\}$, de modo que el semiespacio $\{x \in \Re^p : \omega^\top x + \beta > 0\}$ contiene al conjunto $\{x^u : u \in I, c^u = 1\}$, y el semiespacio $\{x \in \Re^p : \omega^\top x + \beta < 0\}$ contiene al conjunto $\{x^u : u \in I, c^u = -1\}$.

Bajo separabilidad lineal de I, el sistema (2) tiene infinitas soluciones, que generan infinitos hiperplanos distintos. ¿Cómo elegimos una de estas soluciones? La calidad de la clasificación, *sobre la muestra de aprendizaje*, es idéntica: todas clasifican correctamente el 100 % de I. Sin embargo, no todas parecen igualmente razonables. Las Máquinas de Vector de Apoyo se basan en la idea de que es conveniente elegir un hiperplano que esté *alejado* de las dos clases.

Se fija una norma $\|\cdot\|$ en $\Re^p$ para medir las distancias (usualmente la euclídea). Para un objeto $u \in I$, la distancia entre x^u y el semiespacio en el que quedará clasificado incorrectamente viene dada por

$$\rho^u(\omega, \beta) = \text{máx}\left\{\frac{y^u(\omega^\top x^u + \beta)}{\|\omega\|^\circ}, 0\right\}, \tag{3}$$

e.g. [2], donde $\|\cdot\|^\circ$ denota la norma dual a $\|\cdot\|$. Se define el *margen* en la muestra de aprendizaje I como el mínimo ρ^u :

$$\rho^I(\omega, \beta) = \min_{u \in I} \rho^u(\omega, \beta).$$

El clasificador buscado es aquél que no sólo clasifique correctamente a todos los objetos de I, sino que tenga margen máximo.

Geométricamente, la búsqueda del clasificador de máximo margen puede verse como un problema de Localización, [3], pues el problema es equivalente a construir la banda de máxima anchura (las distancias medidas con la norma $\|\cdot\|$) que deja un grupo a cada lado.

Usando la homogenidad de la función margen, el problema de maximización del margen puede ser formulado como el siguiente problema convexo con restricciones lineales:

$$\begin{array}{ll} \text{mín} & \|\omega\|^\circ \\ \text{s.a.:} & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \\ & \omega \in \Re^N, \, \beta \in \Re. \end{array} \tag{4}$$

2.2. El caso no separable

Cuando I no es linealmente separable, el problema (4) es infactible, por lo que deben aplicarse enfoques alternativos. Uno de estos enfoques consiste en aplicar a los datos, como preprocesamiento, una transformación $\phi : \Re^p \to F$, donde F es un espacio vectorial de mayor dimensión (posiblemente infinita), de manera que, en el nuevo espacio, la muestra de aprendizaje $\hat{I} = \{(\phi(x^u), c^u) : u \in I\}$ sea linealmente separable, [6, 9, 10, 11, 15]. Conseguido esto, se buscan $\omega \in F$, $\beta \in \Re$, y se construye la regla de clasificación, que estaría basada en la función f,

$$f(x) = \omega^\top \phi(x) + \beta, \tag{5}$$

que asigna, como es habitual, al grupo 1 si $f(x) > 0$, y al grupo -1 si $f(x) < 0$. Esta regla es lineal sobre los datos transformados, pero no lineal en el espacio original $\Re^p$.

El problema de maximización del margen es

$$\begin{array}{ll} \text{mín} & \|\omega\|^\circ \\ \text{s.a.:} & y^u \left(\omega^\top \phi(x^u) + \beta\right) \geq 1 \; \forall \in I \\ & \omega \in F, \, \beta \in \Re. \end{array} \tag{6}$$

Para el caso en que $\|\cdot\|$ sea poliédrica y F tenga dimensión grande (pero finita), (6) se escribe como un problema lineal de gran tamaño, para cuya resolución son especialmente convenientes técnicas de generación de columnas, permitiendo al mismo tiempo hacer selección automática de variables, [6].

Una estrategia alternativa (y a veces complementaria) para abordar el caso no separable, es la que se basa en la maximización del *márgen débil*, [7, 8, 15], en la que, partiendo del problema infactible (4), se perturban sus restricciones para hacerlo factible, introduciendo una penalización en el objetivo para controlar la perturbación introducida.

Terminamos el análisis comentando que, en una gran variedad de aplicaciones, la importancia del error cometido al clasificar incorrectamente un objeto depende fuertemente del grupo al que éste pertenece: los costes asociados a los falsos positivos y a los falsos negativos pueden ser muy distintos, y, como en el caso del diagnóstico de enfermedades, puede ser difícil cuantificar esa importancia asignando costes. En tal caso podemos plantear el problema biobjetivo de maximización simultánea del margen en cada uno de los dos grupos. Como se prueba en [4], para el caso euclídeo, las soluciones eficientes resultan ser hiperplanos paralelos a la solución del problema de máximo margen clásico (6).

Fijando el ω obtenido al resolver (6), y dejando variar β, se obtienen las distintas soluciones eficientes, que dan distintos niveles de compromiso entre los falsos positivos y los falsos negativos.

3. Método del vecino más cercano

El método del vecino más cercano y sus variantes están basados en la idea intuitiva de que objetos parecidos pertenecen a la misma clase, de manera que la clase a la que pertenece un objeto puede ser inferida a partir de la clase a la que pertenecen los objetos (o el objeto) de la muestra de aprendizaje que más se le parecen.

La idea de parecido es reflejada formalmente en el concepto de *disimilaridad* o de distancia. Dada una disimilaridad $d(\cdot,\cdot)$, la regla de clasificación del vecino más cercano (1-NN), basada en la muestra de aprendizaje I, asignará un objeto $u \in \Omega$ a la clase a la que pertenezca el objeto $v \in I$ más parecido a u, i.e., tal que

$$d(x^u, x^v) = \min_{w \in I} d(x^u, x^w). \tag{7}$$

En caso de empate, debe asignársele una, como por ejemplo la clase más frecuente o una clase elegida aleatoriamente.

Uno de los mayores inconvenientes de este método es que, cuando la muestra de aprendizaje I contiene muchos objetos, requiere una gran cantidad de almacenamiento. Además cada vez que se va a clasificar un nuevo objeto, se necesita calcular su disimilaridad a cada uno de los objetos de I lo que supone un coste computacional muy alto, especialmente si la dimensión p es también alta.

Por ello, han aparecido multitud de métodos en la literatura en los que la muestra de aprendizaje I es sustituida, en la fase de clasificación de nuevos individuos, por otra más pequeña J, a cuyos elementos se les conoce con el nombre de *prototipos*, de manera que el comportamiento de la regla de clasificación no empeore o incluso mejore por ser menos sensible a la presencia de observaciones extrañas o erróneas.

En [3, 5] podemos encontrar una reciente revisión de los métodos basados en prototipos, su formulación como problemas enteros y su resolución exacta o heurística.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por la red temática Análisis y Aplicaciones de Decisiones sobre Localización de Servicios y Problemas Relacionados (MTM2004-22566-E), por el Proyecto BFM2002-04525-C02-02 y por el Grupo de Investigación FQM-329 del Plan Andaluz de Investigación.

Referencias

[1] Apte, C. The big (data) dig. *OR/MS Today*, February 2003.

[2] Carrizosa, E. y Fliege, J. Generalized goal programming: Polynomial methods and applications. *Mathematical Programming*, 93, 281–303, 2002.

[3] Carrizosa, E. y Martín-Barragán, B. Problemas de clasificación: una mirada desde la localización. En *Avances en localización de servicions y sus aplicaciones*. B. Pelegrín (Ed.), pp. 249–276. Servicio de Publicaciones de la Universidad de Murcia, 2005.

[4] Carrizosa, E. y Martín-Barragán, B. Two-group classification via a biobjective margin maximization model Por aparecer en *European Journal of Operational Research*.

[5] Carrizosa, E., Martín-Barragán, B. Romero-Morales, D., y Plastria, F., A Dissimilarity-based Approach for Classification, METEOR Research Memorandum RM/02/027, 2002.

[6] Carrizosa, E., Martín-Barragán, B. y Romero-Morales, M.D., A Biobjective Model to Select Features With Good Classification Quality and Low Cost. *Proceedings of the Fourth IEEE International Conference on Data Mining.* IEEE Publications, 2004. Pag. 339-342.

[7] Cortes, C. y Vapnik, V., Support-vector network. *Machine Learning*, 1, 113–141, 1995.

[8] Cristianini, N. y Shawe-Taylor, J., *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[9] Demiriz, A., Bennett, K.P. y Shawe-Taylor, J., Linear programming boosting via column generation. *Machine Learning*, 46, 225–254, 2002.

[10] Duarte Silva, A.P. y Stam, A., Second order mathematical programming formulations for discriminant analysis. *European Journal of Operational Research*, 72, 4–22, 1994.

[11] Falk, J.E. y Karlov, V.E., Robust separation of finite sets via quadratics. *Computers and Operations Research*, 28, 537–561, 2001.

[12] Fayyad, U. y Uthurusamy, R., Evolving data mining into solutions for insight *Communications of the ACM*, 45:, 28–31, 2002.

[13] Hand, H., Mannila, H. y Smyth, P., *Principles of Data Mining.* MIT Press, 2001.

[14] Hastie, T., Tibshirani, R., y Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Sprnger, 2001.

[15] Herbrich, R., *Learning Theory Classifiers. Theory and Algorithms.* MIT Press, 2002.

[16] Informe de Intel sobre la ley de Moore. `http://www.intel.com/research/silicon/mooreslaw.htm`

[17] Witten, I.H., y Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.